'What's Hard in German? (WHiG): a British learner corpus of German'

Ensslin, A.

Corpora

DOI:
10.3366/cor.2014.0057

Published: 01/11/2014

Publisher's PDF, also known as Version of record

What's Hard in German? WHiG: a British learner corpus of German

Cédric Krummes[1] and Astrid Ensslin[2]

Abstract

This short paper reports on the construction of a freely available learner corpus of advanced British English undergraduate learners of German, which was developed at Bangor University (United Kingdom) and the Humboldt-Universität zu Berlin (Germany). The corpus, What's Hard in German? (WHiG), can be compared to its sibling learner corpus, Falko-L2, and to the native speaker corpus, Falko-L1, using a creative commons user licence or a specifically designed online corpus platform interface, Annis. The main aim of the WHiG project was to collect typed-up essays from participants who achieved a CEFR proficiency of level B2 or higher in the metadata obtained, and whose essays are subsequently POS-tagged, lemmatised and error-tagged. A multi-layer annotation offers researchers the opportunity to work with the collected data and even to provide their own layers of annotation, which in turn can be made available online.

Keywords: argumentative writing, corpus linguistics, error-annotation, German as a Foreign Language, learner corpus, native speaker corpus, student essays

[1] Institut für Germanistik, Universität Leipzig, Beethovenstr. 15, 04107 Leipzig, Germany.
[2] School of Creative Studies and Media, Bangor University, College Road, Bangor, LL57 2DG, United Kingdom.
Correspondence to: Cédric Krummes, *e-mail*: cedric.krummes@uni-leipzig.de

Corpora 2014 Vol. 9 (2): 191–205
DOI: 10.3366/cor.2014.0057

1. Introduction: learner corpora

In this paper, we present a new British learner corpus of German called 'WHiG', named after its project name 'What's Hard in German?'[3] Granger (2002: 7) defines learner corpora as 'electronic collections of authentic [foreign/second language] textual data assembled according to explicit design criteria for a particular [second/foreign language acquisition/teaching] purpose'. For English, for instance, the most enterprising corpus is Granger's International Corpus of Learner English (ICLE), maintained and sustained at the Université Catholique de Louvain, Belgium. Version 2.0 of ICLE is a collection of '3.7 million words of EFL writing from learners representing 16 mother tongue backgrounds' (Faculté de philosophie, arts et lettres, 2010). Two main approaches exist regarding the analysis of learner corpus data: Error Analysis (EA) and Contrastive Interlanguage Analysis (CIA). The former is predominantly associated with Corder (1981) and consists of identifying, quantifying and analysing the errors that are found in learner texts. The latter, coined by Granger (1996), consists of comparing learner language with either native language or with learners with a different first-language background. By comparing these varieties, patterns of over-use, under-use and misuse can be recognised. The aims of learner corpus research are three-fold: (*1*) finding and describing errors and error patterns, (*2*) unveiling the learners' interlanguages, and (*3*) informing and improving learning and teaching materials and methods.

Although English is at the forefront of corpus linguistics and learner corpora, other learner corpora exist, such as the International Corpus of Learner Finnish (ICLF; Jantunen, 2011), the Language Learner Corpus of Norwegian (ASK; Tenfjord, 2004), the French Learner Language Oral Corpora (FLLOC; Myles and Mitchell, 2011) or the Learner's Language Corpus of Japanese (日本語学習者言語コーパス (にほんごがくしゅうしゃげんご); Umino, 2009). The type of learner corpus we are discussing in this paper deals with German as a foreign language, as can be found in the Falko and WHiG corpora (see Section 2).

Currently, a small number of German learner corpora exist (see Goossens and Granger, 2013), and they vary radically in a number of ways: most importantly in terms of learners' first language (L1), skill level and mode of production (written/spoken); and text genre (summaries, essays, *etc*.), compilation purpose (analytical focus), retrieval method, and/or depth and types of annotation. Heike Zinsmeister and Margrit Breckle's AleSKO (Annotiertes Lernersprachenkorpus/'annotated learner language corpus'),

[3] The project 'What's Hard in German?' (WHiG) was co-funded by the Arts and Humanities Research Council (UK) and the Deutsche Forschungsgemeinschaft (Germany), AHRC reference AH/H500081/1. With grateful acknowledgments to Anke Lüdeling and her corpus team at the Humboldt-Universität zu Berlin.

which is based at the Universities of Konstanz, Germany, and Vilnius Pedagogical Institute, Lithuania, looks at learners whose first language is Chinese. The 'Analyzing Discourse Strategies: A Computer Learner Corpus' project led by Christina Frei and Edward Nixon at the University of Pennsylvania focusses on *ab initio* to intermediate-mid learners whose first language is predominantly American English in order to analyse their preferred discourse strategies in threaded discussions, chat and essays; Falko (see Section 2) contains multiple layers of annotation (e.g., error, target hypotheses, POS), covers a wide range of advanced L1s and comprises both summaries and argumentative essays; Andrea Abel and Aivars Glazniek's KOLIPSI (Kompetenzen Linguistiche e Psicosociali/ 'linguistic and psycho-social competences'), which is based at the European Academy Bolzano/Bozen, comprises written output (letters and e-mails) of A2-C1 learners with Italian as their first language; Ulrike Gut's Augsburg-based Learning the Prosody of a foreign language (LeaP) corpus consists of various types of oral learner output, such as read, free and prepared speech, and her learners represent various skill levels. Julie Belz's Telecollaborative Learner Corpus of English and German at Pennsylvania State University was compiled using a very specific retrieval method (telecollaborative partnerships). Finally, Ursula Weinberger's error-tagged learner corpus Corpus of Learner German (CLEG) at Lancaster contains argumentative essays produced by advanced British learners and compiled mostly to analyse matters of modality (Weinberger, 2008). In what follows, we will discuss in detail the aims and scope of WHiG and its parent corpus, Falko.

2. Aims of WHiG and its parent corpus, Falko

Falko stands for 'fehlerannotiertes Lernerkorpus' ('error-annotated learner corpus') and has been designed and compiled at the Humboldt-Universität zu Berlin (see Lüdeling *et al*., 2005) with the following desiderata:

- advanced learners of German (minimum CEFR level B2);
- written texts, typed rather than handwritten; no spoken data;
- error-tagged;
- multi-layered annotation: PoS-tags, target hypotheses, target hypothesis difference markers, macro-structural annotations, verb phrasal annotations are all searchable on their own, in combination with one another, and in combination with token data and meta-data;
- free availability of texts and meta-data;
- versioning of (sub) corpora: keeping old (sub) corpora online when new developments are introduced; and,

- free online availability and, under creative commons licence, free offline availability[4]

Initially, Falko consisted of materials from learners and native speakers who were summarising academic articles. Later on in the project, participants were asked to write argumentative essays instead. This was done for two main reasons: (*1*) to obtain a greater variety of text types and, therefore, to be able to compare text-type effects and register (see Biber, 2009); and (*2*) because learners tend to copy verbatim original passages in summary essays, whereas argumentative essays yield more authentic texts. Due to this change in strategy, Falko now consists of the following subcorpora:[5]

- summary corpus written by learners (version 1.2: 40,865 tokens);
- summary corpus written by native speakers (version 1.2: 21,211 tokens);
- essay corpus written by learners (version 2.0: 132,066 tokens);
- essay corpus written by native speakers (version 2.0: 70,110 tokens);
- essay corpus written by learners at Georgetown University, Washington D.C. (version 1.0: 76,062 tokens); and,
- WHiG essay corpus (version 2.0: 130,187 tokens)

Whereas the essay learner corpus in Falko consists of texts produced by speakers of over forty-nine different first languages (excluding additional languages), the WHiG project, funded from 2009 to 2012, aimed to focus on structural difficulties in British learners of German with a view to informing educational practice at UK higher education institutions in particular. The fieldwork therefore involved collecting data solely from learners of German whose first language is (British) English.[6] The methodology for retrieving learner data, and the corpus design and annotation used for WHiG, was the same as for Falko, in order to ensure comparability of data and consistency of approach.

We give a brief overview of the corpus design in Section 3; Section 4 provides details of the data and metadata collected and how target hypotheses are formed; Section 5 provides an overview of tokens, participants' gender distribution and proficiency levels; Section 6 gives details of the research carried out by team members at Bangor University and Humboldt-Universität

[4] For more details, see: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/zugang/

[5] We acknowledge that the number of tokens in any corpus varies according to the software used and according to the version of that software. In the case of Falko, the numbers are retrieved from the online corpus platform, Annis, (Zeldes *et al.*, 2009).

[6] The exception being native speakers of Welsh: some learners, especially those participants at Welsh universities, were bilingual speakers of Welsh and English, and would indicate Welsh as their first language with English as their other first language or their second language.

zu Berlin; and Section 7 explains the benefits of the corpus and considers possibilities for future research.

3. Corpus design

WHiG (and its parent corpus Falko) was designed according to the desiderata listed above. In order to determine whether participants can be regarded as advanced learners of German, WHiG follows the methodology designed by the parent project, Falko. The meta-data includes a so-called C-test score (see Sigott, 2004) – a cloze test comprising five texts, where each text contains twenty word-final gaps. Each correctly identified gap yields one point, with the final score (between 0 and 100) being assigned a language proficiency level on the Common European Framework of Reference for Languages (CEFR) scale (see Table 5).[7] Furthermore, this method is also used at the Humboldt-Universität zu Berlin to determine the language proficiency levels of incoming non-residential students. Typewritten texts were chosen because of the convenience of not needing to transcribe spoken data but mainly because writing is a skill that is required at advanced levels of language learning. Handwritten texts were avoided to prevent concomitant problems with legibility. The data is error-tagged in order to compare learners' interlanguage with a (hypothesised) target language,[8] while any annotations or tags are added layer-by-layer by use of a stand-off architecture. Differences between the original text and target hypotheses are automatically marked by comparing the different layers of the corpus. This is convenient since it means that searches can be conducted while either looking at only specific annotation levels or by covering two or more search levels simultaneously. A multi-layered annotation system also permits researchers to add their own annotations (in EXMARaLDA or MMAX, for instance) and to carry out their own searches. In order for scholars to conduct their research, WHiG texts are available either through a creative commons licence or through a custom-made online search platform called Annis (Zeldes *et al.*, 2009). Search platforms and their data can undergo changes, which is why the online WHiG corpus is versioned to reflect annotative tweaks. This allows researchers to refer to exact versions of the corpus, thus maximising analytical replicability.

4. Corpus implementation

To retrieve learner data for WHiG, the research team identified and approached partner universities in England and Wales where German was

[7] A distinction does need to be drawn, however, between a C-test score that correlates with a score on the CEFR scale (as is the case in the WHiG/Falko methodolgy) and a CEFR level reached (and/or awarded) after a more comprehensive language assessment.

[8] The tagging was done by three full-time researchers based at Humboldt University, who were regularly cross-checking and discussing their annotative decisions.

offered as a four-year undergraduate degree, which includes a third year spent abroad. Second- and final-year undergraduate students[9] (i.e., immediately before and after the year abroad) would self-select as participants. As WHiG did not record data from the same participants twice[10] but, rather, in terms of cross sections, the corpus can be classified as a *semi*-longitudinal corpus, comparing how progress in German can be achieved.

During the data collection, participants were not asked for personal details such as their names, their date of birth,[11] or their contact details. However, with a view to increasing rates of participation and giving back to the 'community', participants were given the option to request feedback on their essays, for which they had to supply an e-mail address. WHiG participants had to indicate in a meta-data questionnaire their gender, their place and year of study, and list all the modern and ancient languages they knew.[12] They were asked to indicate from what age they had learnt or acquired any of their languages, which of them they considered their 'mother tongue', whether they received instruction (and, if applicable, for how long and in which instructional context), and whether they spent some time abroad in a country where they speak any languages listed (and, if applicable, for how long and where). The final piece of metadata consisted of a C-test score, extracted from participants filling in the gaps in the C-test, for which they were given thirty minutes. The actual data, the argumentative essays, was produced in class under supervision and participants were instructed to type their essay in either Microsoft Notepad or Word[13] whilst seeking no help or support from their peers, online resources, mobile phone resources or staff. Participants were given ninety minutes to write 500 words. Four essay topics, directly taken and translated from the ICLE corpus collection guidelines (Faculté de philosophie, arts et lettres, 2010), were given:

- *Kriminalität zahlt sich nicht aus*. ['Crime does not pay.']
- *Die meisten Universitätsabschlüsse sind nicht praxisorientiert und bereiten die Studenten nicht auf die wirkliche Welt vor. Sie sind deswegen von geringem Wert*. ['Most university degrees are

[9] This was the wording consistently used in WHiG, as partner universities used various words for these two years of study: 'Level 2–Level 3', 'Year 2–Year 3', '2nd Year–4th Year'.
[10] With the exception of five students who have been flagged in the corpus.
[11] In order to comply with the Falko metadata which does require a date of birth, WHiG participants were assigned a fictive date of birth according to their degree level and their academic year. Second-year students were assumed to be nineteen and final-year students were assumed to be twenty-one, which are the typical ages for students at English and Welsh universities.
[12] This included a revision of how the language background could be elicited. Prior to WHiG, Falko participants were asked to draw a schematic diagram from which a meta-data table was created. This was simplified for WHiG (and Falko, thereafter) by rendering the meta-data table questions into a user-friendly and user-tested language background questionnaire.
[13] Microsoft Word was allowed on occasions where no German spell and grammar checkers were installed.

theoretical and do not prepare students for the real world. They are therefore of very little value.']

- *Die finanzielle Entlohnung eines Menschen sollte dem Beitrag entsprechen, den er/sie für die Gesellschaft geleistet hat*. ['A man/woman's financial reward should be commensurate with their contribution to the society they live in.']
- *Der Feminismus hat den Interessen der Frauen mehr geschadet als genützt*. ['Feminism has done more harm to the cause of women than good.']

After three data collection visits, the *Universitätsabschlüsse* ('university degrees') topic was dropped because most participants (perhaps unsurprisingly) had chosen it, thereby threatening to skew the corpus data towards a specific semantic domain. So, for the rest of the WHiG project, participants only had three topics to choose from. Essays were saved as .txt files with the Unicode UTF-8 encoding.

At the Humboldt-Universität zu Berlin, the essays were tagged and lemmatised with the part-of-speech tagger, TreeTagger (Schmid, 1994), using the Stuttgart Tübingen Tag Set (STTS).[14] Later, the essays were error-tagged according to the guidelines set out in Reznicek *et al*. (2012). During the initial period of establishing the guidelines, essays were annotated by two people and an inter-annotated agreement value was calculated. Lüdeling (2008) shows, however, that no two people will annotate ('correct') a learner text in the same way, which is why the project, Falko, with its WHiG data, allows for multiple layers and versions of annotations.

When error-annotating, two target hypothesis layers were added to the raw text (Reznicek *et al*., 2010): ZH1 (*Zielhypothese 1* 'target hypothesis 1') and ZH2 (*Zielhypothese 1* 'target hypothesis 2'). ZH1 rectifies errors concerning orthography, morphology and syntax, whereas ZH2 rectifies errors (or rather misformulations) concerning semantics, lexis, pragmatics and stylistics.[15] A part of the text deemed 'correct' was simply copied into the target hypothesis layers and left untouched. Original text and target hypothesis were also compared and an annotation layer was added describing these changes or differences ('ZH1Diff' and 'ZH2Diff', respectively), as shown in Table 1:

To illustrate both target hypotheses, Figure 1 shows the following utterance from WHiG:[16]

---

[14] For more information, see: http://www.sfs.uni-tuebingen.de/Elwis/stts/Wortlisten/WortFormen.html (accessed 9 January 2013).

[15] Clearly, ZH2 is considerably more prone to subjective annotator judgement and inter-annotator disagreement than ZH1.

[16] This query is retrievable from http://korpling.german.hu-berlin.de/falko-suche/Cite/AQL(%22Schluss%22),CIDS(FalkoEssayL2WHIGv2.0),CLEFT(5),CRIGHT(5) (accessed 9 January 2013), followed by clicking on the 'Show Result' button.

ZH1Diff/ZH2Diff	Meaning	Explanation
INS	insert	a token has been inserted
DEL	delete	a token has been deleted
CHA	change	a token has been changed and stays put
MOVS	move source	a token has been moved away
MOVT	move target	a token has moved to this location
MERGE	merge	two or more tokens have been merged into one
SPLIT	split	one token has been split into two or more tokens

Table 1: Comparing source text and target hypothesis

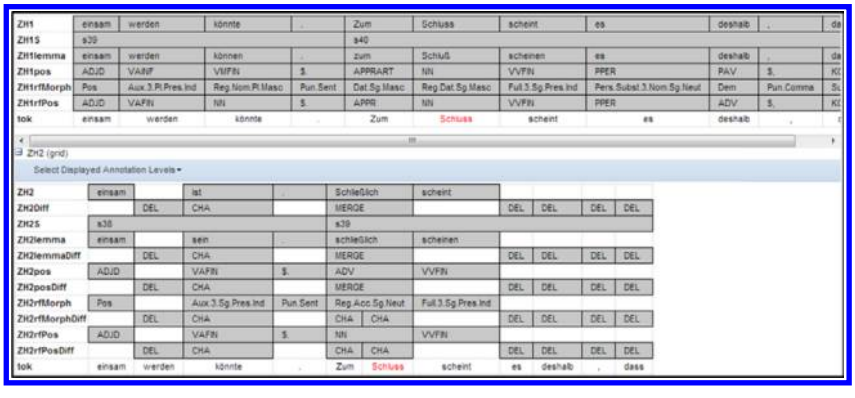

Figure 1: Target hypotheses displayed on Annis (Zeldes *et al*., 2009)

- original and ZH1: *einsam werden könnte*. *Zum Schluss scheint es deshalb*, *dass* [‘could become lonely. At the end it seems therefore that’]
- ZH2: *einsam ist*. *Schließlich scheint* [‘is lonely. Finally it seems’ – rectifying errors of modality and lexis/phraseology]

5. Structure

Between 16 February 2010 and 8 February 2012, 279 essay texts were collected, providing 157,460 tokens. Table 2 summarises how many essays were collected at which university and also shows token counts per university, and the average standardised type/token ratio (STTR). The numbers indicated were calculated in WordSmith Tools 5.0 (Scott, 2008).

University	Texts	Tokens	STTR
Aberystwyth	1	704	71.86
Bangor	29	15,547	69.04
Bristol	27	16,687	73.16
Cambridge	20	11,223	76.36
Lancaster	4	2,582	75.76
Leeds	170	94,475	72.77
QMUL	18	10,113	73.32
Sheffield	2	1,005	77.67
UCL	8	5,124	75.69
Total	279	157,460	
Average no. of tokens per text		564.37	73.96

Table 2: Number of texts and tokens in WHiG (STTR: n=100)

Essay topic	Female		Male		Both	
	n	percent	*n*	percent	*n*	percent
Entlohnung 'remuneration'	27	13.30	11	13.75	38	13.43
Feminismus 'feminism'	83	40.89	16	20.00	99	34.98
Kriminalität 'crime'	51	25.12	34	42.50	85	30.04
Studenten 'students'	42	20.69	19	23.75	61	21.55

Table 3: Essay topic distribution per gender and for both

In terms of gender distribution, eighty male (28.27 percent) and 203 (71.73 percent) female students self-selected as participants. Table 3 provides further details on the gender distribution of participants and the proportion of topics which they chose.

Table 4 presents the distribution of the C-test scores according to year of study and CEFR level. As could be expected, the proficiency levels of second-year students compared with final-year students differed considerably. Carrying out a t-test, there was a significant difference in the C-test scores of second-year (M=74.20, SD=11.46) and final-year students (M=79.31, SD=10.37); $t(281)=2.89$, $p=0.0002$. It is essential to point out, however, that further research is needed to compare the participants' proficiency levels in the C-test with their essay-writing performance.

CEFR level	C-test score range	Second-years		Final-years	
		n	percent	*n*	percent
B1	40–59	9	10.00	9	4.66
B2	60–79	52	57.78	83	43.01
C1	80–89	20	22.22	67	34.72
C2	90–100	9	10.00	34	17.62

Table 4: German proficiency by CEFR level and year of study

6. Analytical research done with the WHiG Corpus

The work of compiling, annotating and managing the corpus is shared between the two partner universities, Bangor University (UK) and the Humboldt-Universität zu Berlin (Germany). These universities have different research priorities and so two analytical strands have emerged in the course of the WHiG project.

At Bangor University, research has concentrated on pragmatic, stylistic and pedagogic aspects of analysis, with particular attention given to formulaic language, collocations and didactic applications. In Jaworska *et al*. (in review), a corpus-driven approach was taken to generate a list of 3-word clusters in both WHiG and the native German corpus, Falko-L1, which is akin to research carried out by Juknevičienė (2009) and Chen and Baker (2010) on clusters found in English essays. Categorising the clusters according to their function (i.e., reference [*in der Vergangenheit* – 'in the past'], essay discourse structure [*in diesem Aufsatz* – 'in this essay'], stance expression [*meiner Meinung nach* – 'in my opinion']), the results showed that 'more types of discourse-structuring devices are found in Falko-L1 (29.56 percent) than in WHiG (23.01 percent), whereas stance expressions are more common in WHiG (24.96 percent) than in Falko-L1 (16.75 percent)' (Jaworska *et al*., in review). A chi-square analysis showed a statistically significant different between the functional distribution between learners and native speakers. Noteworthy in WHiG were the high proportion of discourse-structuring clusters *in diesem Aufsatz werde ich* ['in this essay I will'] and *zum Schluss* ['finally'] followed by a modal verb. These two clusters reflect well the necessity for learners to use boilerplate expressions in their writing; the clusters are, however, not idiomatic in native German. Equally, *wie/als zum Beispiel* ['when/such as for example'] is over-used by British learners; native speakers would have more synonyms at their disposal.

As part of 'giving back to the community' and creating impactful outputs, these findings have fed into a study on using WHiG data to create a learning and teaching resource (Krummes, 2012) documented in Krummes and Ensslin (2012). Although spelling, grammar and lexis presented no

urgent problems for WHiG participants, they did not use much formulaic language and collocations. WHiG data shows that the participants produced inauthentic phrases:

- to introduce the essay topic: *In diesem Aufsatz werde ich über* x *schreiben* ['In this essay I will write about *x*']
- to give examples: *Nehmen wir als Beispiel* x ['Let us take as an example *x*']
- to express a personal stance: *Ich bin der Meinung*, *dass* ['I am of the opinion that']
- to verbalise a conclusion: *Zum Schluss kann man sagen*, *dass* ['At the end one can say that']

Beginner learners of German and near-native speakers, on the other hand, use a higher number of collocations; to alleviate this 'collocational and formulaic dip' (see Biskup, 1992; and Wray, 2002), a worksheet was developed to take into account Webb and Kagimoto's (2011) findings that collocations are best learnt with fewer node words (i.e., keywords) and more examples. In Krummes and Ensslin (2012), we coined the '5–5–5 method': ideally, a worksheet presents five node words (named 'keywords' in the handout) with five formulaic sentences ('phrases') per node word and five concordances ('examples') per phrase. The learning and teaching resource documents five keywords chosen for their ubiquitous use by German scholars and in UK Higher Education German-language settings:

- *Zweck* ['aim']: in introductions (essay discourse structure)
- *Beispiel* and *beispielsweise* ['example']: in the main body (essay discourse structure)
- *Erachtens* and *Ansicht* ['opinion']: in the main body (personal stance expression)
- *laut* and *zufolge* ['according to']: in the main body introductions (impersonal stance expression)
- *Fazit* ['conclusion']: in conclusions (essay discourse structure)

At the Humboldt-Universität zu Berlin, the research focus has primarily been on syntactic issues, such as learners' uses of the German middle field (Reznicek, 2012), underuse of syntactic categories in learner corpora (Hirschmann *et al*., 2011), and learner syntax more generally (Lüdeling, 2012); as well as methodological concerns related to corpus annotation (Golcher and Reznicek, 2011; Lüdeling, 2011; Rehbein *et al*., 2012; Reznicek *et al*., forthcoming; and Reznicek and Krummes, 2011) and analysis (Reznicek and Bennöhr, 2011).

7. Conclusion

The major benefit of the freely available WHiG data is that it allows researchers to discover authentic language patterns that can be used in German language teaching. To this day, German as a foreign language (DaF) materials have largely relied on anecdotal evidence and have not been as corpus-informed as their EFL or ESL counterparts. As already mentioned above, Jaworska *et al*. (in review) and Krummes and Ensslin (2012) have shown that British learners of German rely on non-idiomatic formulaic sequences. This issue has been addressed by producing a corpus-informed worksheet (Krummes, 2012), which has been trialled by thirty-five undergraduate students at Bangor University (Wales) and the University of Leeds (England). Further research could take into account the error-tagging developed at the Humboldt-Universität zu Berlin and analyse anglicisms, for instance, or, taking into account that British learners of German tend to know further foreign languages, gallicisms (e.g., **Epruvetten* instead of *Reagenzgläser* 'test tubes') and hispanisms. Further corpus research may also be able to reveal which phenomena are (still) difficult for advanced learners to learn (e.g., article *versus* zero-article) and which ones may (no longer) be an issue (e.g., *wegen* 'because of'+genitive case). We are confident that the WHiG corpus and its parent project, Falko, are of particular interest to German linguists, corpus linguists and DaF researchers, not least because of the free (online or offline licence) availability and the possibility for researchers to add their own layers of annotations.

References

Biber, D. 2009. 'Multi-dimensional approaches' in A. Lüdeling and M. Kytö (eds) Handbook of Corpus Linguistics, pp. 822–55. Berlin: Walter de Gruyter.

Biskup, D. 1992. 'L1 influence on learners' renderings of English collocations: a Polish/German empirical study' in P.J.L. Arnaud and H. Béjoint (eds) Vocabulary and Applied Linguistics, pp. 85–93. Basingstoke: Macmillan.

Chen, Y.H. and P. Baker. 2010. 'Lexical bundles in L1 and L2 academic writing', Language Learning and Technology 14 (2), pp. 30–49.

Corder, S.P. 1981. Error Analysis and Interlanguage. Oxford: Oxford University Press.

Faculté de philosophie, arts et lettres. 2010. UCL – ICLEv2. Available online, at: http://www.uclouvain.be/en-277586.html

Golcher, F. and M. Reznicek. 2011. 'Stylometry and the interplay of topic and L1 in the different annotation layers in the FALKO corpus', Talk Given at QITL 4. 31 April 2011. Berlin. Accessed 19 March

2012, at: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/vortraege

Goossens, D. and S. Granger. 2013. 'Learner corpora around the world'. Available online, at: http://www.uclouvain.be/en-cecl-lcworld.html

Granger, S. 1996. 'From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora' in K. Aijmer, B. Altenberg and M. Johansson (eds) Languages in Contrast: Text-based Cross-linguistic Studies, pp. 37–51. Lund: Lund University Press.

Granger, S. 2002. 'A bird's-eye view of learner corpus research' in S. Granger, J. Hung and S. Petch-Tyson (eds) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 3–33. Amsterdam: Benjamins.

Hirschmann, H., A. Lüdeling, M. Reznicek, I. Rehbein and A. Zeldes. 2011. 'Underuse of syntactic categories in learner corpora: a case study on modification', Talk Given at Learner Corpus Research (LCR). 17 September 2011. Louvain. Accessed 19 March 2012, at: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/vortraege

Jantunen, J.H. 2011. 'Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi', Lähivõrdlusi. Lähivertailuja 21, pp. 86–105. Available online, at: http://www.rakenduslingvistika.ee/ul/files/LV21_04_Jantunen.pdf

Jaworska, S., C. Krummes and A. Ensslin. In review. 'Formulaic sequences in native and non-native argumentative writing in German: a corpus-driven comparison'. International Journal of Corpus Linguistics.

Juknevičienė, R. 2009. 'Lexical bundles in learner language: Lithuanian learners vs. native speakers', KaLBOTYRa 61 (3), pp. 61–72.

Krummes, C. 2012. 5 Keywords in German Essay-Writing. Accessed 25 November 2013, at: http://www.bangor.ac.uk/creative_industries/documents/WHiG-5KeywordsinGermanEssay-Writing.pdf

Krummes, C. and A. Ensslin. 2012. 'Formulaic language and collocations in German essays: from corpus-driven data to corpus-based materials', Language Learning Journal, pp. 1–18. Available online, at: http://dx.doi.org/10.1080/09571736.2012.694900

Lüdeling, A. 2008. 'Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora' in M. Walter and P. Grommes (eds) Fortgeschrittene Lernervarietäten, pp. 119–40. Tübingen: Niemeyer.

Lüdeling, A. 2011. 'Corpora in linguistics: sampling and annotation' in K. Grandin (ed.) Going Digital: Evolutionary and Revolutionary Aspects of Digitization, pp. 220–43. New York: Science History Publications.

Lüdeling, A. 2012. 'Syntaktische Muster in Texten des Deutschen als Fremdsprache', Talk Given at Vortragsreihe des Sprachwissenschaftlichen Instituts, Ruhr-Universität Bochum. January 2012.

Lüdeling, A., M. Walter, E. Kroymann and P. Adolphs. 2005. 'Multi-level error annotation in learner corpora' in Proceedings of Corpus Linguistics 2005. Birmingham. Available online, at: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/pdf/FALKO-CL2005.pdf

Myles, F. and R. Mitchell. 2011. French Learner Language Oral Corpora. Available online, at: http://www.flloc.soton.ac.uk/

Rehbein, I., H. Hirschmann, A. Lüdeling and M. Reznicek. 2012. 'Better tags give better trees – or do they?', Linguistic Issues in Language Technology 7 (10), pp. 1–18. Accessed 5 June 2014, at: http://elanguage.net/journals/lilt/article/view/2692

Reznicek, M. 2012. 'Lernereffekte im deutschen Mittelfeld', Talk Given at HU Forschungskolloquium. 4 January 2012. Berlin. Accessed 19 March 2012, at: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/vortraege

Reznicek, M. and J. Bennöhr. 2011. 'Korpusanalyse und -auswertung', Psycholinguistischer Methodenworkshop, HU-Berlin. 1 March 2011. Berlin. Accessed 19 March 2012, at: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/vortraege

Reznicek, M. and C. Krummes. 2011. 'Annotation of learner data in the Falko corpus', Workshop given at WHiG Symposium. 18–19 July 2011. Bangor. Accessed 19 March 2012, at: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/vortraege

Reznicek, M., A. Lüdeling and H. Hirschmann. Forthcoming. 'Competing target hypotheses in the Falko Corpus: a flexible multi-layer corpus architecture' in A. Díaz-Negrillo (ed.) Automatic Treatment and Analysis of Learner Corpus Data. Amsterdam: John Benjamins.

Reznicek, M., A. Lüdeling, C. Krummes, F. Schwantuschke, M. Walter, K. Schmidt, H. Hirschmann and A. Torsten. 2012. Das Falko-Handbuch: Korpusaufbau und Annotationen (Version 2.01). Available online, at: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v2.01

Reznicek, M., M. Walter, K. Schmid, A. Lüdeling, H. Hirschmann and C. Krummes. 2010. Das Falko-Handbuch: Korpusaufbau und Annotationen (Version 1.0.1). Accessed 5 June 2014, at: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v1.0.1

Schmid, H. 1994. 'Probabilistic part-of-speech tagging using decision trees' in Proceedings of the International Conference on New Methods in Language Processing, pp. 1–9. Accessed 5 June 2014, at: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

Scott, M. 2008. WordSmith Tools version 5. Liverpool: Lexical Analysis Software. Available online, at: http://www.lexically.net

Sigott, G. 2004. Towards Identifying the C-Test Construct. Bern: Peter Lang.

Tenfjord, K. 2004. 'ASK – a computer learner corpus' in P.J. Henrichsen (ed.) CALL for the Nordic Languages: Tools and Methods for Computer Assisted Language Learning. København: Copenhagen Business School.

Umino, T. 2009. Learner's Language Corpus of Japanese. Accessed 25 November 2013, at: http://cblle.tufs.ac.jp/llc/ja/

Wray, A. 2002. Formulaic Language and the Lexicon. Cambridge: Cambridge University Press.

Webb, S. and E. Kagimoto. 2011. 'Learning collocations: do the number of collocates, position of the node word, and synonymy affect learning?', Applied Linguistics 32 (3), pp. 259–76.

Weinberger, U. 2008. 'Modality as indicator of L2 proficiency? A corpus-based investigation into advanced German Interlanguage' in M. Walter and P. Grommes (eds) Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitspracherwerbsforschung/Advanced Learner Varieties: Corpus Linguistics and Research into Second Language Acquisition, pp. 141–64. Berlin: Niemeyer.

Zeldes, A., J. Ritz, A. Lüdeling and C. Chiarcos. 2009. 'ANNIS: a search tool for multi-layer annotated corpora' in M. Mahlberg, V. González-Díaz and C. Smith (eds) Proceedings of the Corpus Linguistics Conference, CL2009. 20–23 July 2009. University of Liverpool, UK. Available online, at: http://ucrel.lancs.ac.uk/publications/cl2009/358_FullPaper.doc, Corpus platform available online, at: http://korpling.german.hu-berlin.de/falko-suche/search.html